Cornell Data Science Training Program Syllabus

INFO 1998 Introduction to Data Science and Machine Learning

Fall 2017 · Gates G01 · Wednesdays from 5:30 - 6:30 pm

This Course

When you finish this program, you will have the foundation and basic skills to contribute to any subteam in Cornell Data Science. This program introduces various machine learning models, visualization techniques, and data manipulation strategies, with applications in the Python programming language. The program is open to all Cornell students. We look forward to getting to know you over the next eleven weeks!

Course Staff

Lecturers:

- Dae Won Kim (dk444@cornell.edu)
- Jared Junyoung Lim (jl3248@cornell.edu)

TAs:

- Abby Beeler (arb379@cornell.edu)
- Ann Zhang (az275@cornell.edu)
- Cameron Ibrahim (cai29@cornell.edu)
- Kexin Zheng (kz73@cornell.edu)
- Ryan Kannanaikal (rk635@cornell.edu)
- Shubhom Bhattacharya (sb2287@cornell.edu)

Course staff office hours will be announced after the first day of the class.

Course Enrollment

Go to the student center and add INFO 1998, the one with 5-digit course number of 18681. The enrollment cap is 140, so if you want to enroll, do so ASAP!

Pre/Co-requisites

Although there is no pre- or co-requisites specified, we strongly recommend you are either enrolled in **CS 1110** or have a **solid background in Python**. The Course will be taught in Python 3 and we will barely spend time on the basics of programming and Python syntax, so make sure you know them!

Questions In Class

- **General Questions** - If you have a question about the material which is applicable to everyone (nothing like “in my specific project”) please raise your hand and we will clarify the material.
- **Personal Questions** - If you fall behind during lecture for whatever reason, or have questions specific to your own computer/project, please hold your question until after

class, ask it on Piazza, or come to office hours. Unfortunately, this may be quite a large class and we can't give people personalized help during class.
- **Course Content** - If we make a mistake on the board or say the wrong thing - don't raise your hand. Just blurt it out. Don't let us teach something that's incorrect.

Grading

The training program is 1-credit S/U course. Below is the percentage of each assignment:

10% Take-home Quiz 1
10% Take-home Quiz 2
15% Project part A
15% Project part B
15% Project part C
15% Project part D
20% Project part E

If you get above 70% by the end of the semester, you will receive SX, otherwise UX. If you think you will have hard time getting a passing grade, please leave a private post on the Piazza and ask for personal helps.

Assignment Submission

All the quizzes have to be submitted on time. We will deduct 10% off from your quiz grades for each day late.

The project will be done as a group, and each group will have **3 slip days** total throughout the semester. You may choose to use all 3 at once, or divide it up, or not use at all. If you want to use slip days, please let the TAs and instructors know before your submission by posting a private note on the Piazza. If you miss the deadline without slip days, we will deduct 10% off from your project grades for each day late.

Joining the Project Team

This course is not required in order to apply to be a project team member of CDS, and taking this course does not guarantee an acceptance to these sub-teams. If you do well in the course, though, it certainly helps!

Piazza

- If it can be asked on Piazza, ask it on Piazza. We check Piazza as much, if not more, than we check our emails.
- We will not respond to training program questions sent to cornelldatascience@gmail.com unless they are urgent.
- We will not respond to questions sent to our personal emails.

Academic Conduct

All Cornell students are expected to follow the Cornell University Code of Academic Integrity (http://cuinfo.cornell.edu/aic.cfm). Students can consult with the course staffs and other students if they struggle, but all the submissions should be original.

Curriculum

- Week 1 - 9/6 - **Introduction to the Course and Data Science**
  - Course logistics
  - What is data science?
  - Why Jupyter and Python
  - Numpy basics
- Week 2 - 9/13 - **Data Manipulation**
  - Pandas review
  - Introduction to the data pipeline
  - Data manipulation tools and techniques
  - Principal Component Analysis - dimensionality reduction
- Week 3 - 9/20 - **Data Visualization**
  - The importance of visualizing data
  - Basic statistical tools for visualizations
  - Visualizations with matplotlib
  - Interactive visualization with plotly*
- Week 4 - 9/27 - **Basics of Machine Learning**
  - Train/Test Frameworks
  - Introduction to supervised learning
  - Functional vs matrix representation
  - Goals of regression
  - Introduction to linear regression
- Week 5 - 10/4 - **Introduction to Classifiers**
  - Extensions of linear regression
  - Loss functions
  - Classifiers basics
  - Decision boundary
  - The concept of sensitivity and specificity
  - ROC curves
- Week 6 - 10/11 - **Application of Supervised Learning: Classifier part 1**
  - Introduction to SVM
  - Hard vs soft margin
  - Use of kernels in machine learning
- Week 7 - 10/18 - **Application of Supervised Learning: Classifier part 2**
  - Introduction to logistic regression
  - Conditional probability
  - Introduction to decision tree
- Week 8 - 10/25 - **Clustering and Unsupervised Learning**
  - Basics of unsupervised learning
  - Latent variables
  - Recommendation systems and filtering techniques

    - Hierarchical clustering and dendrograms
    - K-means clustering
    - Principal Component Analysis revisited
- Week 9 - 11/1 - **Bias-Variance Trade Off**
    - Risk and loss functions
    - Bias and variance
    - The tradeoff between bias and variance
    - Underfitting and overfitting
    - Feature/Subset Selection techniques
- Week 10 - 11/8 - **Model Selection and Optimization**
    - Model selection
    - Hyperparameters
    - Regularization Methods
    - Validations
- Week 11 - 11/15 - **Meta-Learning and ensembles**
    - Recap of the bias-variance tradeoff
    - Introduction to ensembles
    - Bagging, Boosting, and Stacking